Probability Assessment with Maximum Entropy in Bayesian Networks

Wim Wiegerinck Tom Heskes
SNN, University of Nijmegen,
Geert Grooteplein 21, 6525 EZ, Nijmegen, The Netherlands
{wimw,tom}@mbfys.kun.nl

Abstract

Bayesian networks are widely accepted as tools for probabilistic modeling. In building Bayesian networks in collaboration with domain experts, the definition of the graphical structure is usually relatively easy. The assessment of the conditional probability tables (CPT) is often a much more difficult task, even when there is a lot of statistical information available as domain knowledge. The problem is that in many cases it is not possible to fill this information directly into the CPTs. In this paper we propose a method to fit the CPTs such that the model reproduces the available information as accurate as possible. We will discuss some criteria to do this and we illustrate the methods in an example.

1. Introduction Probabilistic graphical models, and in particular Bayesian networks, are nowadays well established as a modeling tool for expert systems in domains with uncertainty [1, 2, 3, 4]. The reason is that graphical models provide a powerful and conceptual transparent representation for probabilistic models. Their graphical structure, showing the conditional independencies between variables, allows for an easy interpretation. On the other hand, since a graphical model uniquely defines a joint probability model, the mathematical consistency and correctness are guaranteed. In other words, there are no assumptions made in the methodology. All assumptions in the model are contained in the definition of variables, the graphical structure of the model, and the parameters in the model.

The construction of a Bayesian network consists of two parts, a qualitative and a quantitative part. The qualitative part is the determination of the structure of the network. It is common practice that structure of the network is determined by hand, especially when the network is built in collaboration with domain experts. Luckily, the determination of this structure is often a relatively straightforward task, since it usually fits well with knowledge that the domain expert has about causal relationships between variables. The quantitative part consists of specifying the conditional probability tables (CPTs) in the network. When there is insufficient data to fully determine the CPTs, these CPTs need also be set by, or in collaboration with, the domain experts [5]. This is considered a much harder or even impossible

task [6, 7], especially since the domain experts very often have no intuition about these probabilities. Luckily, in many cases statistical information $\mathcal{I}$ is available, e.g. from the literature. In such a case, one can try to choose the CPTs in the network such that the probability model induced by the CPTs complies with this information $\mathcal{I}$. Unfortunately, $\mathcal{I}$ often does not translate directly into network CPTs, that is to say, it is often not clear to the experts how $\mathcal{I}$ should be translated into quantitative CPTs in the Bayesian network. This is, for example, the case when conditional probabilities are given in the 'wrong direction', from 'effect' to 'cause'.

The availability of additional statistical information has been noted before. In [7] the authors propose to exploit this information by expressing them as constraints on models. Next, they sample the models that satisfy these constraints. In this paper, we propose an alternative method to translate $\mathcal{I}$ to the CPTs. Our method is different in the sense that it yields a single 'best' model instead of a number of reasonable models. In many applications, we expect a single model to be much more convenient to work with, while still being sufficiently accurate given the available information. Another difference is that, in our approach, the constraints need not be exactly satisfied. In case of constraint violation, our method yields a model that minimizes the 'violation amount'.

In this paper, we will assume that $\mathcal{I}$ can be formulated as conditional probabilities of the form $Pr(U_\alpha|v_\alpha) = Q_\alpha(U_\alpha|v_\alpha)$. A typical example is the sensitivity of a medical test. This is the the probability of a positive test result U_α given that the patient has disease v_α. An obvious requirement for the model P is to reproduce these probabilities Q_α as accurate as possible. To measure 'violation amount' or inaccuracy, we need an error or distance function between the given conditional probabilities Q_α and the conditional probability according to the graphical model P. A natural distance between a model distribution and a target distribution is the Kullback-Leibler divergence. Minimizing this distance, however, is in general not sufficient for a unique determination of the model P. A whole manifold of models with the same distance to $\mathcal{I}$ may exist. A standard way to proceed is to select a representative of this manifold of distributions through the Maximum Entropy method (MaxEnt) [8]. MaxEnt searches for the distribution that maximizes entropy under the given constraints. Roughly speaking, it selects the distribution P that satisfies the constraints without introducing any additional information. An alternative is to assume a prior distribution P_0 and to try to use the freedom in the parameters of the model to stay as close as possible to P_0.

This paper is organized as follows. In section 2, we propose our procedure and we discuss the various choices that can be made. In section 3, the method is applied on a toy problem. We end the paper with a short discussion in section 4.

2. Parameter fitting Our starting point is a parameterized probabilistic model $P(X|\theta)$ with parameters θ on a set of discrete variables $X = X_1, \ldots, X_n$ in a finite domain, $X_i \in \{1, \ldots, n_i\}$. In a standard Bayesian network, $P(X) = \prod_i P(X_i|\pi_i)$, the parameters θ are the probability tables, $\theta = \{\theta_{X_i,\pi_i} = P(X_i|\pi_i)\}$.

Now we want to find parameters θ^* such that the probabilistic model reproduces a given set $\mathcal{I}$ of conditional probabilities

$$Pr(U_\alpha|v_\alpha) = Q_\alpha(U_\alpha|v_\alpha) \qquad \alpha = 1 \ldots m \tag{1}$$

as accurate as possible. Here, v_α is a setting of states of variables V_α for which the conditional probability of the variables U_α is given in the table Q_α. For example $v_1 = (X_2 = \mathit{true}, X_6 = \mathit{false})$, $v_2 = (X_2 = \mathit{true}, X_6 = \mathit{false})$ etc. If $U_1 = (X_3, X_4)$, then the conditional distribution $Q_1(X_3, X_4|X_2 = \mathit{true}, X_6 = \mathit{false})$ is given, etc. We do not assume that the different distributions in set $\mathcal{I}$ are mutually consistent. So obviously, there need not to be a parameter setting θ^* such that $P(U_\alpha|v_\alpha, \theta) = Q_\alpha(U_\alpha|v_\alpha)$ for all α. The best one can do is to fit P to $\{Q_\alpha\}$, using some distance or error function between probability distributions.

A natural choice for a distance function between the model P and a target Q_α is the Kullback-Leibler divergence,

$$KL(Q_\alpha(U_\alpha|v_\alpha)||P(U_\alpha|v_\alpha, \theta)) \equiv \sum_{\{u_\alpha\}} Q(u_\alpha|v_\alpha) \log \frac{Q(u_\alpha|v_\alpha)}{P(U_\alpha|v_\alpha)} \,, \tag{2}$$

where the sum over $\{u_\alpha\}$ stands for the sum over all states of U_α. The Kullback-Leibler divergence has the nice properties that it is nonnegative and equal to zero if and only if $P(U_\alpha|v_\alpha, \theta) = Q_\alpha(U_\alpha|v_\alpha)$. To fit all conditional probabilities, we propose to minimize the cost function

$$F(\theta) = \sum_\alpha \rho_\alpha KL(Q_\alpha(U_\alpha|v_\alpha)||P(U_\alpha|v_\alpha, \theta)) \,, \tag{3}$$

where one might want to consider non-equal weights $\rho_\alpha > 0$ if one considers some conditional distributions more important than other ones. One particular and quite natural choice for ρ_α follows by choosing a probability distribution $R(x)$ and setting

$$F(\theta) = \sum_{\{x\}} R(x) \sum_\alpha \delta_{x v_\alpha} KL(Q_\alpha(U_\alpha|v_\alpha)||P(U_\alpha|v_\alpha, \theta)) \,, \tag{4}$$

in which $\delta_{x v_\alpha} = 1$ if the states x and v_α are compatible, and $\delta_{x v_\alpha} = 0$ otherwise.

Minimizing $F(\theta)$ does not provide a unique set of parameters. To see this consider the factorized model of two variables $P(x_1, x_2) = P(x_1)P(x_2)$. If the given set $\mathcal{I}$ consists of one distribution $Q(x_1)$, then all parameters such that $P(x|\theta) = Q(x_1)P(x_2|\theta)$ are minimizers of $F(\theta)$. For a practical solution, a single representative would be desirable. For this we need an additional cost function $G(\theta)$. This additional cost function is only needed to get rid of the freedom in θ that still remains after minimizing F. In other words, the optimal parameters are given by

$$\theta^* = \operatorname*{argmin}_\theta \lim_{c \to \infty} [F(\theta) + \frac{1}{c} G(\theta)] \,. \tag{5}$$

Several choices can be made for $G(\theta)$. We will discuss two of them. The first one is the maximum entropy choice (relative to some distribution P_0, e.g. the constant distribution),

$$G_{Entr}(\theta) = KL(P(x)||P_0(x)) \,. \tag{6}$$

This choice is motivated by the maximum entropy (MaxEnt) principle. This principle states that, under the constraints that θ is a minimizer of $G(\theta)$, the model $P(x|\theta)$ should contain no additional information [relative to $P_0(x)$]. The second choice interprets $P_0(x)$ as a prior guess for the model. Here the idea is that the solution should 'fit' $P_0(x)$ under the constraints that θ is a minimizer of $F(\theta)$. This leads to

$$G_{Prior}(\theta) = KL(P_0(x)||P(x)) \; . \tag{7}$$

Which G to choose is, in our opinion, mainly a matter of taste: the results in practice will be roughly the same (see also below). In both interpretations, taking the same reference distributions for weighing the constraints $[R(x)]$ and minimizing the remaining degrees of freedom $[P_0(x)]$, seems a reasonable standard choice.

In our experience, the best way to minimize $[F(\theta) + \frac{1}{c}G_{Entr}(\theta)]$ for a large value of c, is through an 'annealing' procedure. Starting from a random θ, we minimize the cost function for $c = c_1$, with c_1 a small value, yielding an optimum $\theta^*_{c_1}$. In the next step, we minimize the error for a larger $c = c_2 > c_1$, starting from the previous optimum $\theta^*_{c_1}$. This procedure is iterated for $c = (c_1 < c_2 < \ldots < c_n)$ until the optimum at the final $c = c_n$ value is obtained. This 'annealing' process of 'cooling down' $\frac{1}{c}$ to zero is crucial in order to find a minimum of $G(\theta)$ in the manifold of parameters that minimizes $F(\theta)$. This annealing technique is often used in constraint optimization using a penalty function [9].

3. An example: coronary heart disease We illustrate our method by an example involving the diagnosis of coronary heart disease, taken from [10]. It is well known that the prior probability of a patient having coronary heart disease depends on sex and age of the patient. Older men have a much higher probability of having the disease than young women. A typical symptom of this disease is *Angina Pectoris* (chest pain). We will construct a model on the basis of information given in [10], which is a biannual publication by the Dutch national health insurances, and which is distributed among all Dutch physicians.

In this example, we have four variables: *age* (a), *sex* (s), *heart-disease* (d), and *chest-pain* (c) . In the example, *age* has four states (30-39, 40-49, 50-59, 60-69), *sex* has two states (*male*, *female*), *heart-disease* has two states (*true*, *false*), and *chest-pain* has four states, (asymptomatic, non-AP pain, atypical AP-pain, typical AP-pain). Assuming that *sex* and *age* are independent, and that the type of *chest-pain* is independent of *sex* and *age* given that we know the state of *heart-disease*, we build a graphical structure according to figure 1.

In [10] the conditional probabilities $Q(d|a, s, c)$ are tabulated (see table 1). Furthermore we assume that s and a are homogeneously distributed (although this is not stated in [10]). For simplicity, we put this assumption in the tables $P(a) = 0.25$, $P(s) = 0.5$ and keep them fixed throughout the optimization process.

In the first model [referred to as model (1)], we assume no other information except $Q(d|a, s, c)$. Therefore we choose the maximum entropy approach, in which both $R(x)$ and $P_0(x)$ are a flat distribution. We minimized $[F(\theta) + \frac{1}{c}G_{Entr}(\theta)]$ for $c = 1$ to $c = 10^6$, where at each iteration we increased c with a factor 10. With slower annealing we obtained the same results, but without annealing we did not

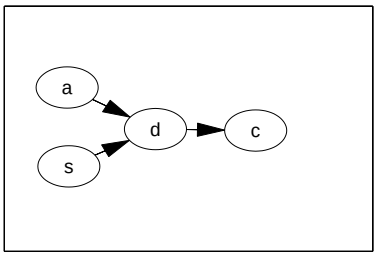

FIG. 1. *Structure of the network for coronary heart disease with four variables:* age *(a),* sex *(s),* heart-disease *(d), and* chest-pain *(c)*

manage to achieve comparable performance. The CPTs $P(d|a,s)$ and $P(c|d)$ that follow from this MaxEnt approach are given in tables 2 and 3 in the rows labeled with model (1). With these CPTs, the model reproduces the desired conditional probabilities with a maximal absolute difference of $\max|Q(d|a,s,c)-P(d|a,s,c)| = 0.007$ and a maximal relative difference of

$$\max\left|\frac{Q(d|a,s,c)-P(d|a,s,c)}{Q(d|a,s,c)}\right| = 0.07\,.$$

The summed Kullback-Leibler divergence for model (1) is $F(\theta) = 5.4\times10^{-4}$.

In the second model [model (2)], we minimized $[F(\theta)+\frac{1}{c}G_{Prior}(\theta)]$ with the same parameter settings as above. We obtain about the same errors relative to Q. The resulting CPTs $P(d|a,s)$ and $P(c|d)$ are given in tables 2 and 3 in the rows labeled with model (2). The marginal distributions that are obtained in this way are slightly closer than to the (flat) marginals of the prior $P_0(x)$ than for model (1). This is in particular the case for $P(c)$, see table 5.

In both models, the CPTs $P(d|a,s)$ indicate a quite high (and worrying) a priori probability of having the heart disease: 35 percent. Furthermore, the model suggests that about 70 percent of all healthy people have some complaints of chest pain (non-AP pain, atypical AP pain, typical AP pain), and that 6 percent of the healthy people have even typical AP pain. In other words, the likelihood ratio $P(c|d)/P(c|\neg d)$ of chest pain is quite low for severe pain (see table 4). If we consider the whole population, the model suggests that all states of chest pain are about equally probable.

This unappealing behavior is most easily explained when we interpret $P_0(x)$ as our prior. This prior states that the a priori probability (i.e. prevalence) of having a disease is 50 percent, and that the a priori distribution of the state of chest pain is a homogeneous distribution over all states of pain. In the next model [model (3)] we assume that we still do not know the prevalence, but we guess that it is 2 percent, so we state $P_0(d) = 0.02$. Since this is a prior, we prefer the method with G_{Prior}. The results are tabulated in tables 2 and 3 in the rows labeled with model (3). The errors relative to Q in this model are, again, not significantly different from the previous models. But now the prevalence of the disease is reduced to the more reassuring value of 13 percent. The probability of an individual to have chest pain is much

sex	age	asympt	non-AP	atyp-AP	typ-AP
m	30-39	0.019	0.052	0.218	0.677
m	40-49	0.055	0.141	0.461	0.873
m	50-59	0.097	0.215	0.589	0.92
m	60-69	0.123	0.281	0.671	0.943
f	30-39	0.003	0.008	0.042	0.258
f	40-49	0.01	0.028	0.133	0.552
f	50-59	0.032	0.084	0.324	0.794
f	60-69	0.075	0.186	0.544	0.906

TABLE 1

Conditional probabilities of heart disease given age, sex and type of chest-pain (asymptomatic, non-AP pain, atypical AP-pain, typical AP-pain). This table is taken from the literature and serves as a constraint for the probability model in figure 1.

more skewed towards asymptomatic (no complaints), see table 5. Furthermore, the probability that a healthy person has considerable chest pain is now reduced to about 6 percent, and even less than one percent for severe chest pain (see table 3).

Next, in model (4), we treat the knowledge that among the Dutch patients the prevalence for this disease is 2 percent [10] as a fact, rather than as an assumption. That is, in the optimization we enforce both $Q(d|a,s,c)$ from table 1 and $Q(d) = 0.02$. We take the weightings ρ_α proportional to the inverse of the number of parent states in Q, so $\rho_\alpha = 1/32$ for $Q(d|a,s,c)$ and $\rho_\alpha = 1$ for $Q(d) = 0.02$.

Apart from these target distributions, we have no further information. Therefore we choose the MaxEnt cost function $G_{Entr}(x)$. The results are tabulated in tables 2 and 3 in rows labeled with model (4). Since this model tries to approximate two (conditional) distributions [$Q(d|a,s,c)$ and $Q(d)$], it is not surprising that the error is larger than in the previous models. The maximal error is $\max|Q(d|a,s,c) - P(d|a,s,c)| = 0.065$ and the maximal relative difference is

$$\max\left|\frac{Q(d|a,s,c) - P(d|a,s,c)}{Q(d|a,s,c)}\right| = 0.6\,.$$

The summed Kullback-Leibler divergence with respect to $Q(d|a,s,c)$ is 0.1. This model predicts a prevalence $P(d) = 0.027$, which is in reasonable agreement with the target. The probability of an individual having any type of chest pain is now reduced to a mere 3 percent (see table 5). The probability that a healthy person has no chest pain at all is 97 percent, and the probability that a healthy person has considerable chest pain is now reduced to less than 0.1 percent: in this model it is exceptional that a healthy person has severe chest pain typical for this disease (see 3). The worrying number in this model is $p(c = asympt|d) = 0.83$, which means that if you have a heart disease, you will probably not notice it. For that reason, we would prefer model (3), that treats the information in table 1 about $Q(d|a,s,c)$ as 'hard evidence', yet the statement $Q(d) = 0.02$ as 'soft evidence'.

4. Conclusion It is still common practice to do the qualitative and quantitative construction of Bayesian networks in collaboration with domain experts.

Model	m/f	30-39	40-49	50-59	60-69
(1)	male	0.19	0.43	0.56	0.64
(1)	female	0.04	0.12	0.30	0.51
(2)	male	0.19	0.42	0.55	0.64
(2)	female	0.04	0.12	0.29	0.50
(3)	male	0.05	0.15	0.24	0.30
(3)	female	0.01	0.03	0.09	0.20
(4)	male	0.01	0.03	0.05	0.07
(4)	female	0.001	0.005	0.02	0.04

TABLE 2

Conditional probabilities (percentages) of heart disease ($d = \text{true}$) conditioned on age and both sexes. These CPTs are obtained by optimizations in various models. See the text for a description of the models.

Model	d	asympt.	non AP	atypical AP	typical AP
(1)	*true*	0.03	0.07	0.33	0.57
(1)	*false*	0.31	0.33	0.30	0.06
(2)	*true*	0.03	0.07	0.27	0.63
(2)	*false*	0.37	0.33	0.23	0.07
(3)	*true*	0.22	0.25	0.26	0.27
(3)	*false*	0.66	0.27	0.056	0.007
(4)	*true*	0.84	0.13	0.022	0.007
(4)	*false*	0.97	0.02	0.0008	0.00003

TABLE 3

Conditional probabilities (percentages) of having a certain state of chest pain (asymptomatic, non-AP pain, atypical AP-pain, typical AP-pain), given the state of heart disease (true or false). These CPTs are obtained by optimizations in various models. See the text for a description of the models.

Model	asympt.	non AP	atypical AP	typical AP
(1)	0.08	0.22	1.1	9.0
(2)	0.08	0.22	1.2	9.3
(3)	0.33	0.90	4.7	38
(4)	0.86	5.3	29	231

TABLE 4

Likelihood ratio $P(c|d)/P(c|\neg d)$ of the probability to have a certain state of chest pain (asymptomatic, non-AP pain, atypical AP-pain, typical AP-pain), given the state of having heart disease divided by the probability of having the same pain, given that there is no heart disease. The ratio is tabulated for different models. See the text for a description of the models.

Model	asympt.	non AP	atypical AP	typical AP
(1)	0.21	0.24	0.31	0.24
(2)	0.26	0.24	0.24	0.26
(3)	0.60	0.27	0.08	0.04
(4)	0.97	0.03	0.001	0.0002

TABLE 5

Probability $P(c)$ to have a certain state of chest pain (asymptomatic, non-AP pain, atypical AP-pain, typical AP-pain), averaged over the whole population. The probability is tabulated for different models. See the text for a description of the models.

The qualitative part, which is the construction of the graph of the model, is usually relatively straightforward. The quantitative part, which is the determination of the conditional probability tables, is often much more difficult for domain experts. However, if other statistical information about the domain is available, in particular if this information is available in terms of a set of conditional probabilities that cannot directly be used as parameters of the model, then fitting this information might be a useful and practical method for setting the parameters of the model. In this paper we proposed a cost function, which consists of two terms. The first, most important one, is used to minimize the distance between the model and the available information, given as 'hard evidence'. The second one is used to select a representative of models that are at equal distance from this information. In this second term one might include a prior guess for the model, i.e., more 'soft evidence'. If one is reluctant to give any information, one can choose the maximum entropy term, which tries to exclude any bias in the model.

Another method for the quantitative assessment of CPTs has been proposed in [7]. An advantage of this method is that it can also handle inequality statements between conditional probabilities. In any case, however, the method assumes that the model can exactly satisfy all statistical information. This, unfortunately, might not the case in practice, due to a too simple model, or even due to inconsistencies in the given information. Another advantage of our method is that it is based on the minimization of a cost function, rather than on a type of rejection sampling. Cost function methods are usually much faster than sampling methods. One of our objectives is therefore to test our method on larger problem domains. Challenges range from the derivation of (approximate) algorithms for efficient minimization to the design of software for automatic incorporation of all kinds of information.

Acknowledgments This project is funded by the Dutch Technology Foundation STW. We thank Jan Neijt for pointing us at the medical example.

REFERENCES

[1] J. Pearl. *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Inc., 1988.

[2] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Statistical society B*, 50:154–227, 1988.

[3] F.V. Jensen. *An introduction to Bayesian networks.* UCL Press, 1996.

[4] E. Castillo, J. M. Gutierrez, and A. S. Hadi. *Expert Systems and Probabilistic Network Models.* Springer, 1997.

[5] H. Wang and M.J. Druzdzel. User interface tools for navigation in conditional probability tables and elicitation of probabilities in Bayesian networks. In *Proceedings of the Sixteenth Annual Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, 2000. Morgan Kaufmann Publishers.

[6] M.J. Druzdzel and L.C. van der Gaag. Building probabilistic networks: "where do the numbers come from?" Guest editors introduction. *IEEE Transactions on Knowledge and Data Engineering*, 12:481–486, 2000.

[7] M.J. Druzdzel and L.C. van der Gaag. Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 141–148, 1995.

[8] R. Levine and M. Tribus, editors. *The Maximum Entropy Formalism.* 1979.

[9] D. Luenberger. *Linear and Nonlinear Programming.* Addison-Wesley, Reading, Massachusetts, 1984.

[10] H.A.I.M. van Leusden, editor. *Diagnostisch Kompas.* CVZ, Amstelveen, NL, 1999.